Royal Institute of Technology

MACHINE LEARNING 2 - THE EM ALGORITHM CONT.

LAST LECTURE

- ⋆ Slicing
- ⋆ Sampling
- ⋆ K-means (inspiration)
- ⋆ EM for GMM

$$p(Z_t = k, Z_{t+1} = l \,|\, x_{1:T})$$

THIS LECTURE

- ★ EM algorithm for GMM
- ★ Baum-Welch - EM algorithm for training an HMM
- ★ Intuitive derivation of EM

RELATIONS BETWEEN LOG-LIKELIHOODS AND Q-TERMS

log-likelihood

Q-term or expected complete log-likelihood (ECLL)

$$Q(\theta, \theta^i) = \sum_n E_{p(Z_n|x_n,\theta^i)}\left[l(\theta; Z_n, x_n)\right]$$

Theorem: by increasing the ECLL (Q-term), we increase the likelihood.

The ECLL may not increase in every step!

EM-ALGORITHM IN GENERAL

- E-step: compute $E_{p(Z_n|x_n,\theta^i)}\left[l(\theta; Z_n, x_n)\right]$
- M-Step:

$$\theta^i = \text{argmax}_\theta \sum_n E_{p(Z_n|x_n,\theta^i)}\left[l(\theta; Z_n, x_n)\right]$$

- Stop when solution or likelihood hardly change otherwise repeat

- ★ Starting points
- ★ Number of starting points
- ★ Sieving starting points
- ★ The competition
  - The first iterations of EM show huge improvement in the likelihood. These are then followed by many iterations that slowly increase the likelihood. Conjugate gradient shows the opposite behaviour.

PRACTICAL ISSUES

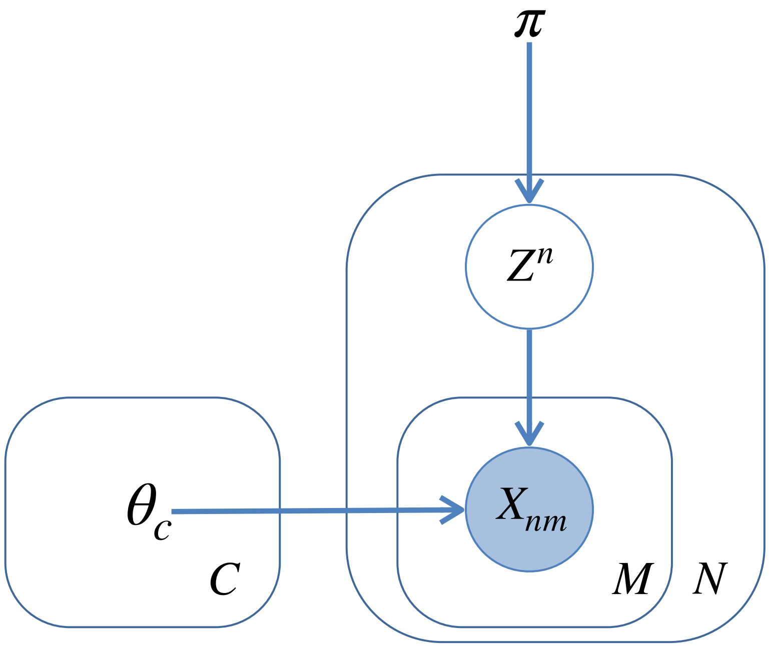
GAUSSIAN MIXTURE MODEL

$\mathcal{D} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ Each a vector

$p(Z = c) = \pi_c$

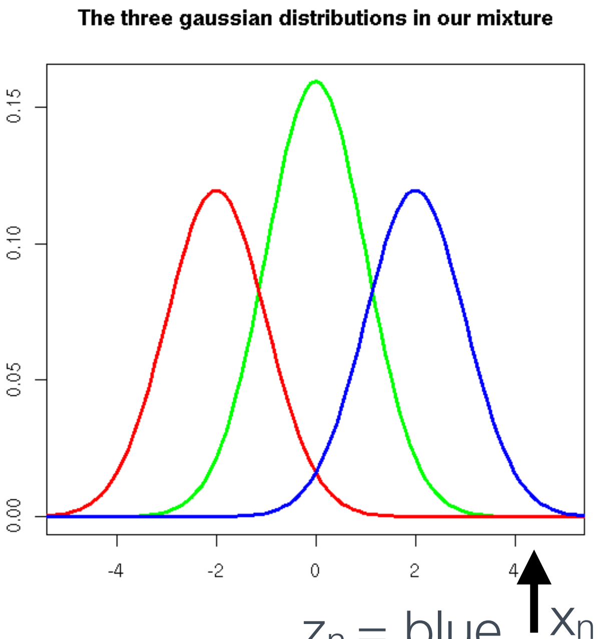
$p(X \,|\, Z = c) = \mathcal{N}(X \,|\, \mu_c, \sigma_c)$

$\boldsymbol{\theta}_c = (\boldsymbol{\mu}_c, \sigma_c)$



EXAMPLE

z_n is red with probability 1/2, green with probability 3/10, blue with probability 1/5



x_n

z_n = blue

z_n is generated as above

x_n is generated from the Gaussian indicated by z_n

We get $x_1, \ldots, x_N$

GMM EM-ALGORITHM

- E-step: compute $r_{nc} = p(Z_n = c|x_n, \theta^i)$
- M-Step: maximize (1) mixture coefficients and (2) each

$$\sum_n r_{nc} \log \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{1}{2\sigma_c^2}(x_n - \mu_c)^2\right)$$

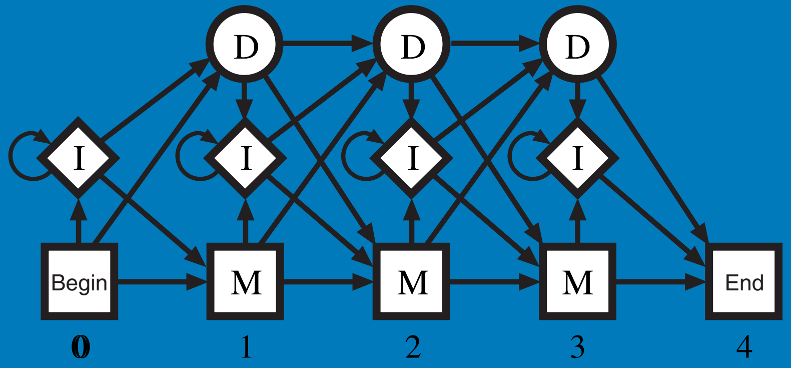
by setting

$$\mu_c = \frac{\sum_n r_{nc} x_n}{r_c}$$ and $$\sigma_c^2 = \frac{1}{\tau_c^2} = \sum_n r_{nc}(x_n - \mu_c)^2 / r_c$$

- set $\theta^{i+1} = \theta$
- Stop when solution or likelihood hardly change otherwise repeat

PROBLEMS ON THREE LEVELS

	x	x	.	.	.	x
bat	A	G	-	-	-	C
rat	A	-	A	G	-	C
cat	A	G	-	A	A	-
gnat	-	-	A	A	A	C
goat	A	G	-	-	-	C
	1	2	.	.	.	3



- Probability of data: $p(x_{1:T})$
- Viterbi (MAP)

  argmax $p(z_{1:T}|x_{1:T})$
- Posterior samples:

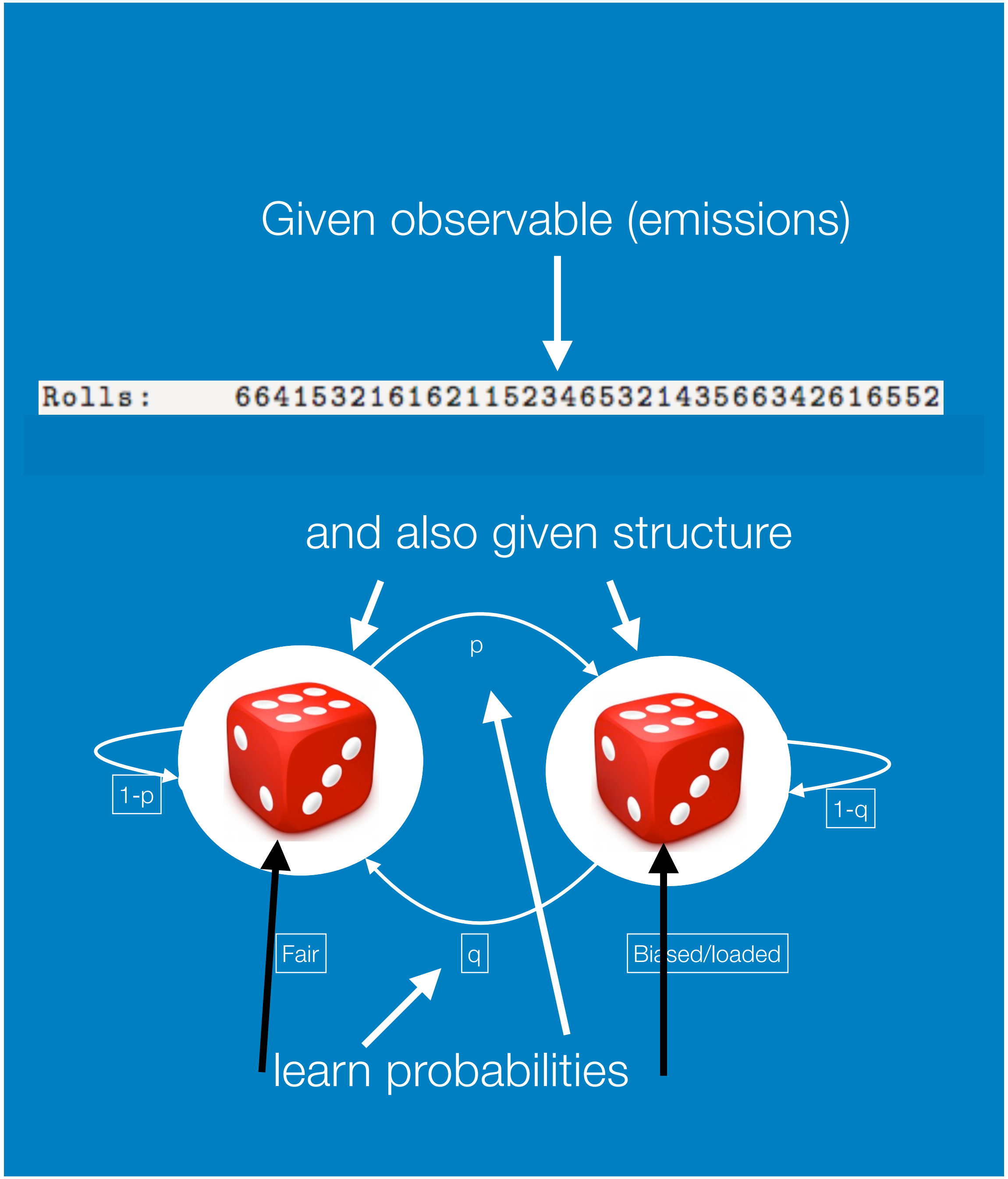
  $\sim p(z_{1:T}|x_{1:T})$

- Parameters:

  given D & struct.
- Structure and param.:

  given D

BAUM-WELCH: LEARNING HMM PARAMETERS


Given observable (emissions)

```
Rolls:     6641532161621152346532143566342616552
```

and also given structure

learn probabilities

- ★ Starts in the state z_1
- ★ When in state z_t
  - outputs $p(x_t|z_t)$ B_{x_t,z_t}
  - moves to $p(z_{t+1}|z_t)$ A_{z_{t+1},z_t}
- ★ Stops after a fixed number of steps or when reaching a stop step

The parameters we now want to learn

LEARNING TRANSITION AND EMISSION PARAMETERS - FULLY OBSERVED DATA

- ★ Parameters
  - transition $A_{lk} = p(Z_t = l | Z_{t-1} = k)$
  - emission $B_{sk} = p(X_t = s | Z_t = k)$

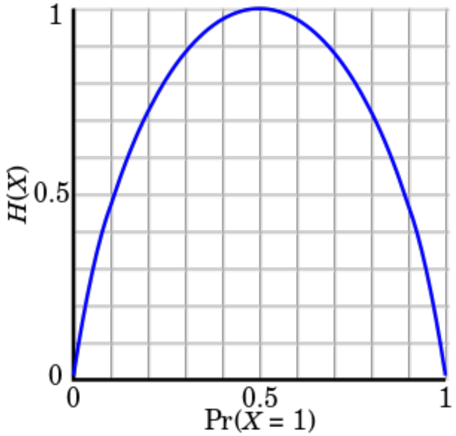
- ★ Data

$$\mathcal{D} = \{(x^1_{1:T}, z^1_{1:T+1}), \ldots, (x^N_{1:T}, z^N_{1:T+1})\}$$

USEFUL INFORMATION THEORY CONCEPTS

Entropy

$$\mathrm{H}(p) = -\int p(x)\log p(x)\mathrm{d}x$$



Use Kullback-Leibler (KL) “distance”

$$\mathrm{KL}(p\,||\,q) = \int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}x$$

THE PUMP

$$\log p(X\,|\,\theta^i) = \mathscr{L}(q(Z),\theta^i) + \mathrm{KL}(q(Z)\,||\,p(Z\,|\,X,\theta^i))$$

THE PUMP

Although negative

$$\log p(X|\theta^i) = \mathscr{L}(q(Z),\theta^i) + \mathrm{KL}(q(Z)||p(Z|X,\theta^i))$$

THE PUMP

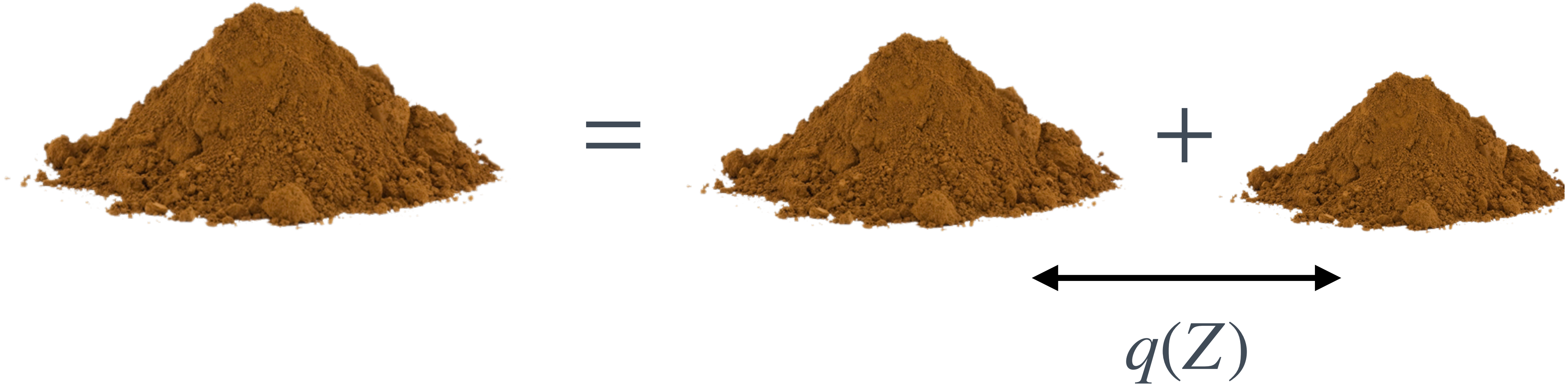
= +

$$\log p(X \,|\, \theta^i) = \mathscr{L}(q(Z), \theta^i) + \mathrm{KL}(q(Z) \,||\, p(Z \,|\, X, \theta^i))$$

THE PUMP

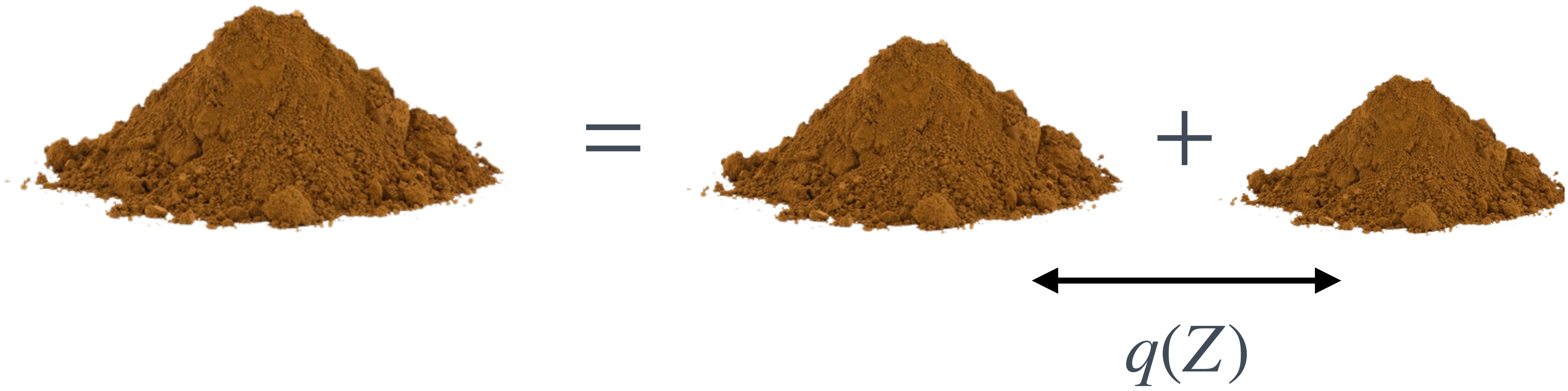
$$\log p(X\,|\,\theta^i) = \mathscr{L}(q(Z),\theta^i) + \mathrm{KL}(q(Z)\,||\,p(Z\,|\,X,\theta^i))$$

THE PUMP



$$\log p(X \mid \theta^i) = \mathscr{L}(q(Z), \theta^i) + \mathrm{KL}(q(Z) \mid\mid p(Z \mid X, \theta^i))$$

THE PUMP



$$\log p(X|\theta^i) = \mathscr{L}(q(Z),\theta^i) + \mathrm{KL}(q(Z)||p(Z|X,\theta^i))$$

$q(z) = p(Z|X,\theta^i)$ makes become 0

THE PUMP

=

$\mathscr{L}(q(Z), \theta^i)$ ←

$q(Z) = p(Z | X, \theta^i)$

Hardwire θ^i in $q(Z)$

$$\log p(X | \theta^i) = \mathscr{L}(q(Z), \theta^i)$$

THE PUMP

$$\mathscr{L}(q(Z), \theta^i) \longleftarrow q(Z) = p(Z|X, \theta^i)$$

$$\log p(X|\theta^i) = \mathscr{L}(q(Z), \theta^i)$$

For any θ,

$$\log p(X|\theta) = \mathscr{L}(q(Z), \theta) + \mathrm{KL}(q(Z)||p(Z|X, \theta))$$

THE PUMP

=

$\mathscr{L}(q(Z), \theta^i)$ $\longleftarrow$ $q(Z) = p(Z \mid X, \theta^i)$

$$\log p(X \mid \theta^i) = \mathscr{L}(q(Z), \theta^i)$$

For max θ $\wedge\vert$ $\Leftarrow$ $\wedge\vert$

$$\log p(X \mid \theta) = \mathscr{L}(q(Z), \theta) + \mathrm{KL}(q(Z) \,||\, p(Z \mid X, \theta))$$

THE PUMP

=

$$\mathscr{L}(q(Z),\theta^i) \longleftarrow q(Z) = p(Z|X,\theta^i)$$

Depends on $q(z)$

$$\mathscr{L}(q(Z),\theta) =$$

$$E_{q(Z)}[\log \frac{p(X,Z|\theta)}{q(Z)}] = E_{q(Z)}[\log p(X,Z|\theta)] + C$$

THE PUMP

$$\mathscr{L}(q(Z), \theta^i) \longleftarrow \boxed{q(Z) = p(Z|X, \theta^i)}$$

Depends on $q(z)$

$$\mathscr{L}(q(Z), \theta) =$$

$$E_{q(Z)}[\log \frac{p(X, Z|\theta)}{q(Z)}] = E_{\boxed{q(Z)}}[\log p(X, Z|\theta)] + C$$

THE PUMP

- Initialize θ^0
- Iterate
  - Min KL by setting $q(Z) = p(Z|x, \theta^i)$ so $\log p(x|\theta^i) = \mathcal{L}(q(Z), \theta^i)$
  - Max $\mathcal{L}(q(Z), \theta)$ w.r.t the θ, notice $\theta^i$is “locked” in $q(Z)$, i.e., ECLL
    - $p(x|\theta) > p(x|\theta^i)$ and KL may increase in the eq for this newθ
  - Set θ^{i+1} to θ

THE END